Unsupervised Correspondence and Co-Segmentation in Latent Diffusion Models - Proof of Concept

Department of Physics, LUMS

Rehan Ahmad

Code | Rehan Ahmad | rehan.ahmad0900@gmail.com | +92 3316658335

1 Introduction

Recent advancements in diffusion models have demonstrated the power of self-attention and cross-attention in uncovering meaningful object structures within images. Previous works have leveraged self-attention to perform segmentation by exploiting its ability to group spatially related regions without requiring explicit supervision [8]. Similarly, cross-attention maps have been utilized to establish local correspondences between images [6] and find keypoints in an unsupervised fashion [5].

These methods are training-free or require only minimal training as they directly exploit the attention maps of pre-trained diffusion models. However, they suffer from key deficiencies. While [8] introduces an efficient way to cluster Self-Attention using a grid-based approach, the method relies upon hyper-parameters such as the merging threshold and the number of points on the grid to obtain a natural clustering of meaningful objects in the image such as "eye", "paw" and "ears". Similarly, while [7] is able to systematically train an embedding to focus on an object of interest, the position that the embedding must focus on must be defined manually before training. More crucially, the method is limited to "local correspondences".

Given these constraints, we ask: *Is it possible to leverage the supervisory signals within self-attention to extract key features across a collection of objects—without manually defining positions or relying on hand-tuned hyperparameters*? Towards this aim, we perform feature matching between two images, I_s and I_t using SD2 to find semantically consistent regions across images without requiring explicit training. This follows the recent trend of using the over-parametrized pipeline of Diffusion Models for extracting useful representations [4] [9].

2 Method

Given a feature map f^l extracted from a U-Net layer , self-attention computes a similarity-based transformation:

$$A_{self} = (Q^l, K^l) = \text{softmax}\left(\frac{Q^l (K^l)^T}{\sqrt{d}}\right) \tag{2.1}$$

Where $Q \in \mathbb{R}^{H \times D \times N}$ represent the query matrix, H is the number of attention heads, D is the feature dimension per head, and N is the number of spatial tokens. Similarly, $K \in \mathbb{R}^{H \times D \times N}$ denote

the key matrix. The dimensions of A_{self} are $A_{\text{self}} \in \mathbb{R}^{H \times N \times N}$. For our purposes, we average all the attention maps we are working with across heads:

$$A = \frac{1}{H} \sum_{h=1}^{H} A^{(h)} \in \mathbb{R}^{N \times N} \tag{2.2}$$

Where $N \in \{64, 256, 1024, 4096\}$. As proof of our concept, we specifically take the middle layer of U-NeT where $N = 1024$ in our notebook. We partition the self-attention matrices A as $\sqrt{N} \times \sqrt{N} \times \sqrt{N} \times \sqrt{N}$. Thus A is partitioned into N matrices. A submatrix can be denoted as $A[i, j, x, y]$ where (i, j) are the pixels of interest and $(x, y) \in \mathbb{R}^{\sqrt{N} \times \sqrt{N}}$ shows the associations of each pixel with the other. Each submatrix $A_{x,y}$ associated with (i, j) is normalized along its rows to ensure that it constitutes a valid probability distribution:

$$A_{i,j} = \frac{A_{i,j}}{\sum_{k=1}^{N} A_{i,j}(k, :)}, \quad A_{i,j} \in \mathbb{R}^{\sqrt{N} \times \sqrt{N}}. \tag{2.3}$$

To look for semantic correspondences and perform feature matching across images I_s and I_t, we utilize **cross self-attention**. Also known as mutual self-attention or KV injection, cross-self-attention have been extensively used in Stable Diffusion II to perform wide-array of tasks such as Style Transfer [3], Localized Edits [2] and Appearance Transfer [1]. In Cross Self-Attention features f^l_{ref} from a source image I_{source} interact with the features from a target image I_{target}. Thus, the usual Self-Attention matrix is replaced as following:

$$A_{cs} = \left(Q^l_{\text{source}}, K^l_{\text{target}}\right) = \text{softmax}\left(\frac{Q^l_{\text{source}} \, (K^l_{target})^T}{\sqrt{d}}\right) \tag{2.4}$$

Where,

$$Q^l_{target} = W_Q f^l_{target} \quad K^l_{source} = W_k f^l_{source}$$

We define two types of cross-self attention maps:

- A cross-self attention matrix that captures how source queries attend to target queries.
- Another that captures how target queries attend to source queries.

After averaging across heads and normalizing, these maps are expressed as:

$$A_{\text{stg}} = \text{softmax}\left(\frac{Q_s \cdot K_t^T}{\sqrt{D}}\right), \quad A_{\text{stg}} \in \mathbb{R}^{N \times N} \tag{2.5}$$

$$A_{\text{tts}} = \text{softmax}\left(\frac{Q_t \cdot K_s^T}{\sqrt{D}}\right), \quad A_{\text{tts}} \in \mathbb{R}^{N \times N} \tag{2.6}$$

Our method follows a principle akin to inverse consistency in image registration which quantifies the error in bidirectional transformations. In this spirit, we define an **inverse mapping score** to evaluate how well a transformation from a source domain to a target domain preserves spatial relationships when mapped back to the source domain.

Given a mapping $f : S \to T$ that transforms a source coordinate x_s into a target coordinate x_t, the inverse mapping function $g : T \to S$ attempts to recover x'_s from x_t. The reconstruction error is defined as:

$$d(x_s, x'_s) = \|x_s - g(f(x_s))\| \tag{2.7}$$

where x_s is the original point in the **source domain**, $x_t = f(x_s)$ is the corresponding mapped point in the **target domain**, and $x'_s = g(x_t)$ is the reconstructed source point from the target mapping. Using cross-self attention, we define $d(x_s, x'_s)$ as following:

$$IMS = d(x_s, x'_s) = \|x_s - A_{tts}(A_{stt}(x_s))\| \tag{2.8}$$

We observe that certain points in the target image receive disproportionately high attention, as reflected in their mapping density. These points indicate feature collapse, where the attention mechanism overly focuses on specific locations. To quantify this, we use **mapping density** which involves counting how many times a target pixel appears during the transformation $A_{stt}(x_s)$.

$$MD(y) = \sum 1(Tgt \to Src = y) \tag{2.9}$$

where 1 is an indicator function that counts occurrences of y. High $MD(y)$ indicates too many inputs mapping to one point and low $MD(y)$ indicates that the target is sparsely mapped.Figure 1 illustrates the mapping density distribution for a sample target image I_t

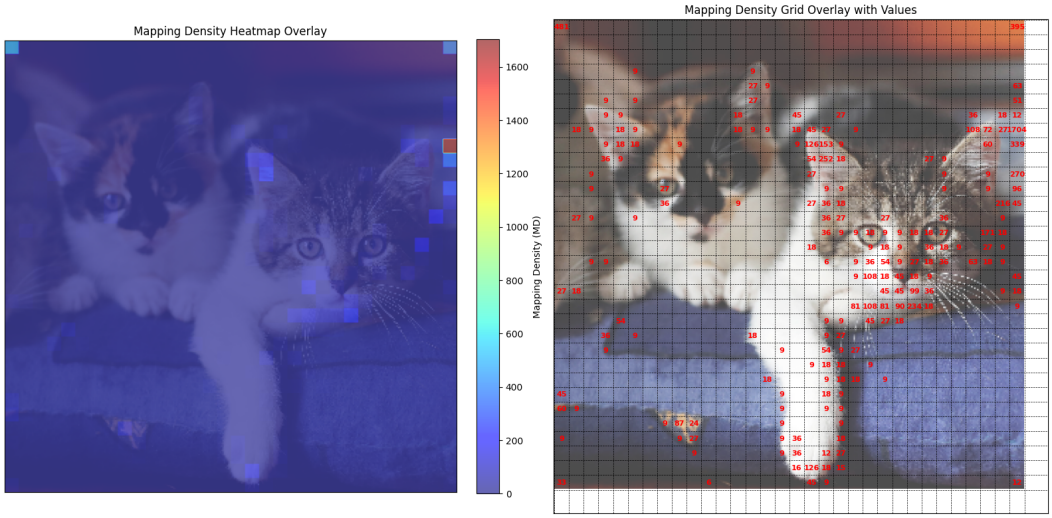

Figure 1: Mapping Density Visualization: High-density areas highlight feature collapse, where attention disproportionately focuses on specific regions in the target image.

A more nuanced quantity to quantify the quality of mappings is **Inverse mapping entropy**. This is used to quantify the uncertainty in the inverse mapping process. Given a probability distribution $P(S_i|T_j)$ that represents the likelihood of a source coordinate S_i mapping back from a target coordinate T_j, the Inverse Mapping Entropy is computed as:

$$H(T_j) = -\sum_i P(S_i|T_j) \log P(S_i|T_j) \tag{2.10}$$

where $P(S_i|T_j)$ is the probability of source location S_i being the correct inverse mapping of target location T_j. If a target point T_j maps back to a single, well-defined source S_i with high probability (i.e., $P(S_i|T_j) \approx 1$), the entropy is low which indicates that the mapping is deterministic. If a target

point T_j maps back to multiple sources S_i with nearly equal probabilities, the entropy is high and indicate the mapping is uncertain.

Our preliminary findings seem to suggest that low IME values are found in the background and smooth regions. On the other hand, regions of interest such as facial features and edges seem to have high IME but not as high as collapse points. Thus, we see that feature of interest exhibit moderate IME.

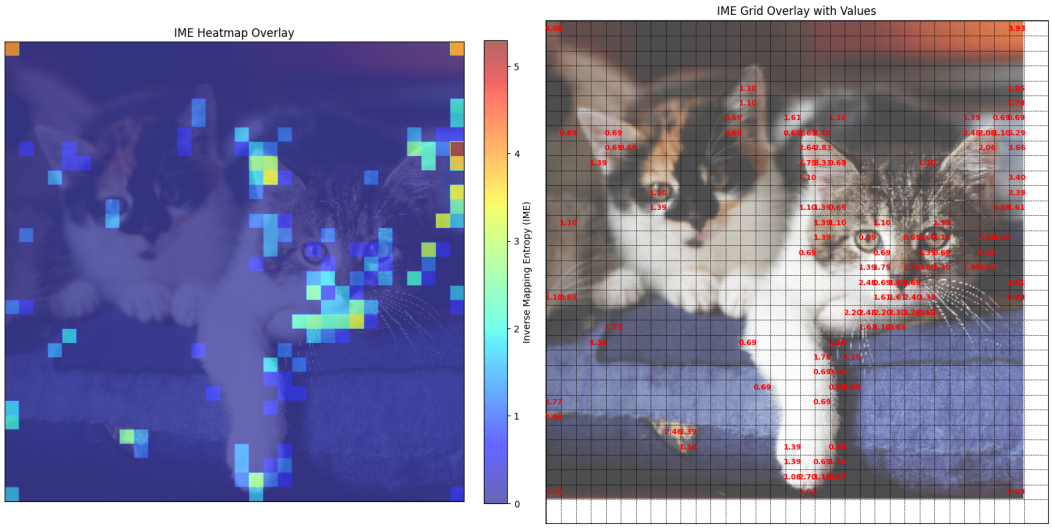

Figure 2: Inverse Mapping Entropy Visualization: High entropy areas correspond to regions with greater mapping uncertainty, while low entropy areas have deterministic inverse mappings.

Building on our discussion of IME, mapping density, and inverse mapping entropy, we propose a method to identify semantically significant regions by balancing these metrics. Rather than considering only the mapping of a single point (i, j) in the target image, we define a connected neighborhood around the point and analyze the distribution of mapped points within this local region:

$$N(i,j) = \{(i + \Delta i, j + \Delta j) \mid \Delta i, \Delta j \in \{-1, 0, 1\}\} \tag{2.11}$$

We then pass this distribution through A_{sst}. For each of the point, we identify the **most-attended pixel** in the target image:

$$(i', j') = \text{argmax}(A_{stg}(i, j) \tag{2.12}$$

For each of the target correspondences (i', j'), we extracts the target-to-source attention slice:

$$(i'', j'') = \text{argmax}(A_{stg}(i', j') \tag{2.13}$$

This approach ensures that local spatial context is incorporated and therefore allows us to distinguish quality keypoints from collapse points. Specifically, if a neighborhood collapses onto a single point, the keypoint is problematic. On the other hand, if the neighbors are mapped in proximity to the keypoint, the mapping is likely well-structured and of high quality.

After analyzing the statistical distribution of Inverse Mapping Entropy and Mapping Spatial Error across the dataset, we establish empirical thresholds to ensure the selection of reliable keypoints. Specifically, we retain keypoints where the IME is below 2 and where the MSE falls within approxi-

mately 20.

To improve upon our estimates, we perform a second pass. This time, we take the target image as the source and vice versa, and carry out the same steps as above. Finally, we track the mapping density between a source and target image. We find that our method is able to capture regions of higher significance as these regions receive more correspondences. For example, in the figure 3, we see that cat's face and eyes receive more correspondences. This indicate they are key features in the mapping process. Edges of the cat bodies receive moderate hit counts. On the other hand, background regions have low-or-zerop mapping.

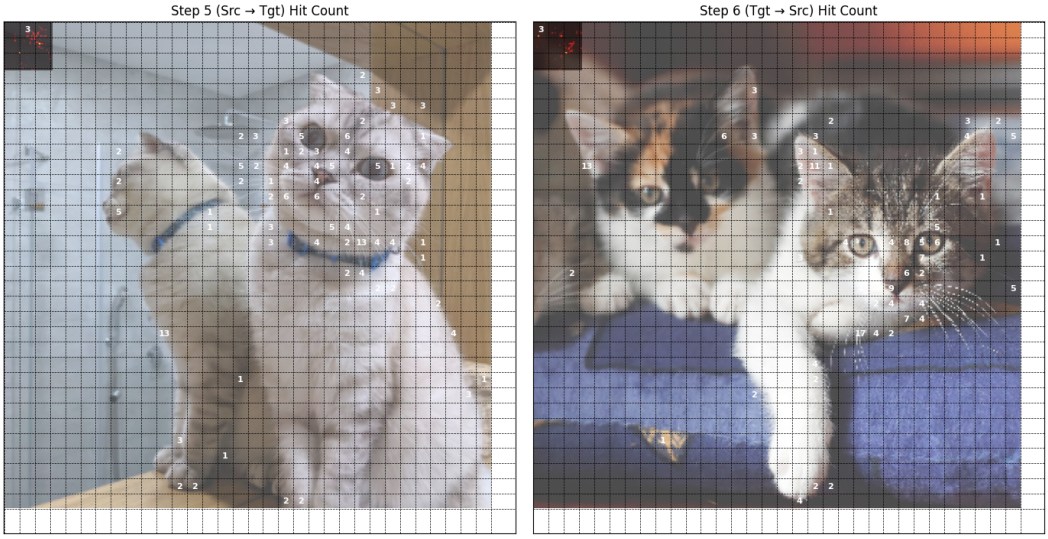

Figure 3: The left image represents the hit count when mapping from the source to the target (Src → Tgt), while the right image shows the hit count when mapping from the target to the source (Tgt → Src).

Once we have these regions, we can perform clustering to remove noise. For example, we see that we get some hits for background in 3 but these are far away from clustering points. End of the page contains some of the visualizations for particular regions that we get from our method. Our method so far is laid down in the following notebook. Moving forward, we aim to use metrics such as KL-divergence and other distance measures between the attention maps of the clustered points in the source and target image to perform co-segmentation. To extend our method and test it more widely across an extensive dataset, we will use the CUB-200 dataset. This is a widely recognized benchmark for fine-grained image classification and object detection tasks. It consists of approximately 6,033 images (around 30 images per class). For our purposes, we will select a particular species and make the evaluate our methods through quantifiable metrics.

The raw link of the Jupyter notebook is the following:
https://github.com/RehanAhmad13/Diffusion-Models/blob/main/Proof_of_concept.ipynb

References

[1] Yuval Alaluf, Daniel Garibi, Or Patashnik, Hadar Averbuch-Elor, and Daniel Cohen-Or. Cross-image attention for zero-shot appearance transfer. In ACM SIGGRAPH 2024 Conference Papers, pages 1–12, 2024.

[2] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 22560–22570, 2023.

[3] Sooyeon Go, Kyungmook Choi, Minjung Shin, and Youngjung Uh. Eye-for-an-eye: Appearance transfer with semantic correspondence in diffusion models. arXiv preprint arXiv:2406.07008, 2024.

[4] Yutong He, Naoki Murata, Chieh-Hsin Lai, Yuhta Takida, Toshimitsu Uesaka, Dongjun Kim, Wei-Hsiang Liao, Yuki Mitsufuji, J Zico Kolter, Ruslan Salakhutdinov, et al. Manifold preserving guided diffusion. arXiv preprint arXiv:2311.16424, 2023.

[5] Eric Hedlin, Gopal Sharma, Shweta Mahajan, Xingzhe He, Hossam Isack, Abhishek Kar, Helge Rhodin, Andrea Tagliasacchi, and Kwang Moo Yi. Unsupervised keypoints from pretrained diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 22820–22830, 2024.

[6] Eric Hedlin, Gopal Sharma, Shweta Mahajan, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. Unsupervised semantic correspondence using stable diffusion. Advances in Neural Information Processing Systems, 36, 2024.

[7] Eric Hedlin, Gopal Sharma, Shweta Mahajan, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. Unsupervised semantic correspondence using stable diffusion. Advances in Neural Information Processing Systems, 36, 2024.

[8] Junjiao Tian, Lavisha Aggarwal, Andrea Colaco, Zsolt Kira, and Mar Gonzalez-Franco. Diffuse attend and segment: Unsupervised zero-shot segmentation using stable diffusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3554–3563, 2024.

[9] Yingheng Wang, Yair Schiff, Aaron Gokaslan, Weishen Pan, Fei Wang, Christopher De Sa, and Volodymyr Kuleshov. Infodiffusion: Representation learning using information maximizing diffusion models. In International Conference on Machine Learning, pages 36336–36354. PMLR, 2023.

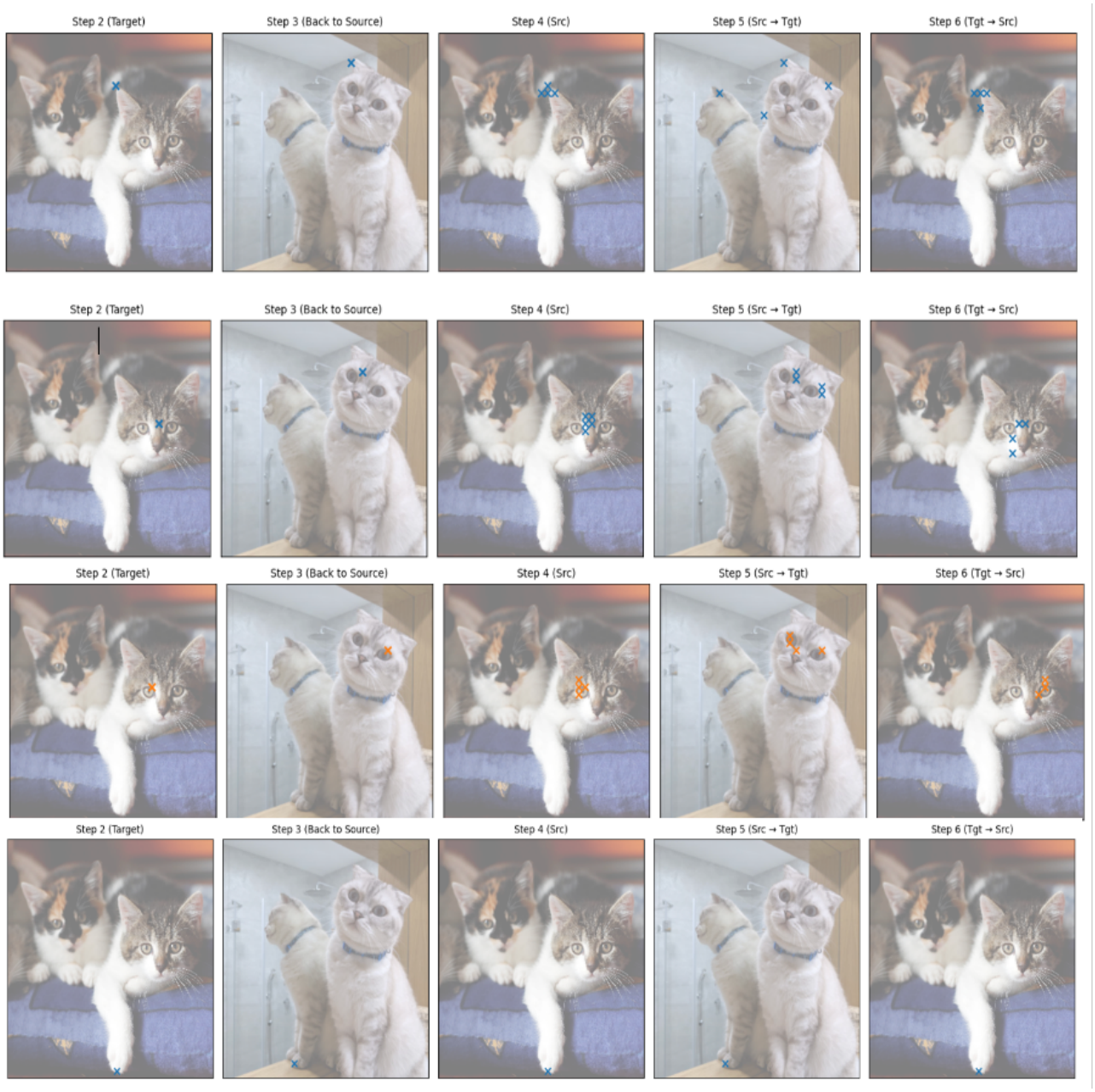

Figure 4: Visualization of hit counts for correspondences across different body parts. Each row represents a different anatomical region: the first row corresponds to the ear, the second to the nose, the third to the eye, and the last to the feet. The left column shows the mapping from source to target (Src $\rightarrow$ Tgt), while the right column shows the mapping from target to source (Tgt $\rightarrow$ Src).